Achieving sustainable nanomaterial design though strategic cultivation of big data

Achieving sustainable nanomaterial design though strategic cultivation of big data

Standardization and interoperability of data for both functional and environmental performance properties of nanomaterials is essential to accelerate sustainable design

Desirée L. Plata[1] and Nina Z. Janković[1,2]

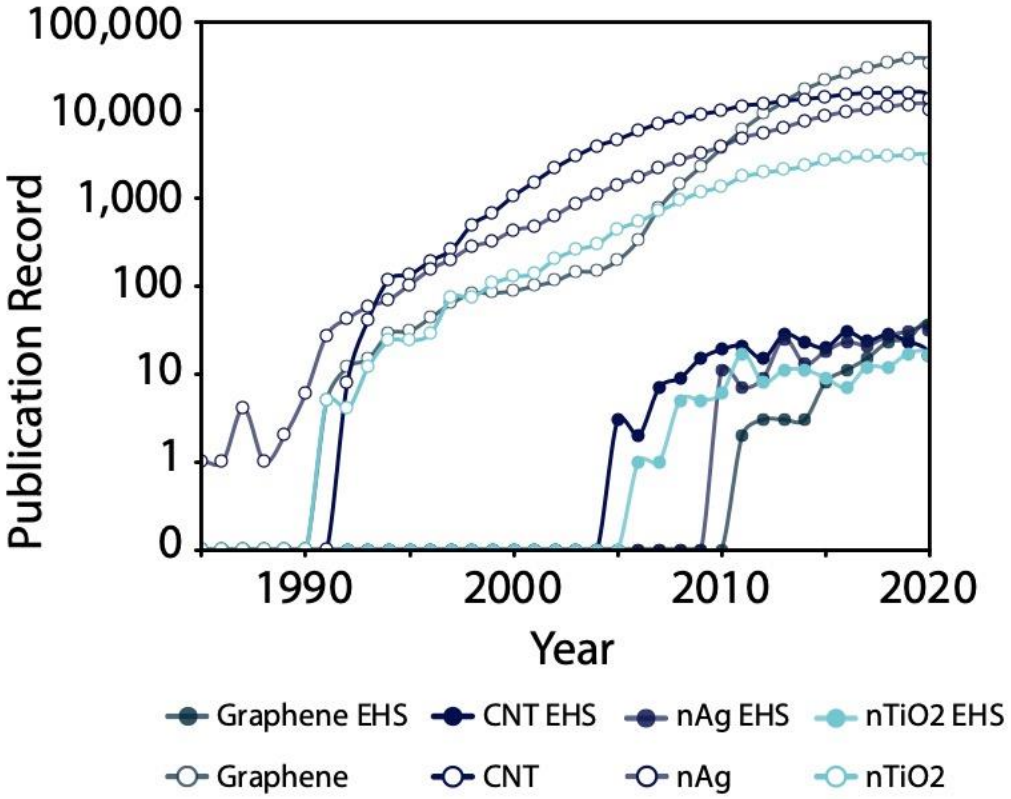
[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge MA 01742

[2]Department of Chemical and Environmental Engineering, Yale University, New Haven CT 06511

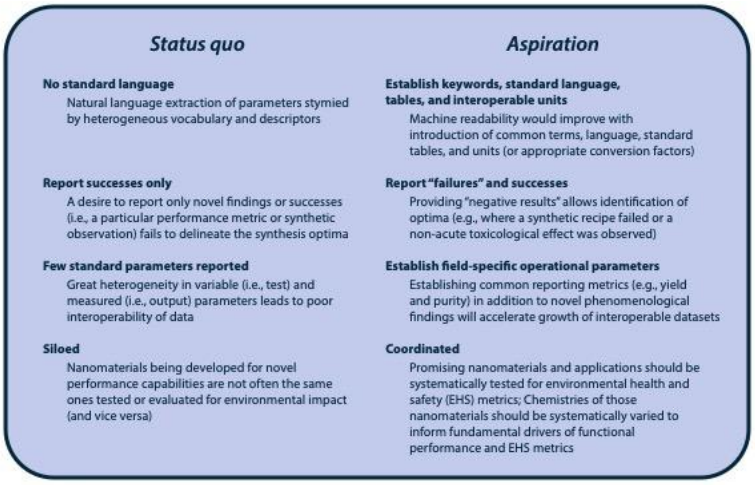
Environmental sustainability is one of the most technically challenging and societally concerning issues of our time. For many today, it is clear that we should use resources more efficiently, that chemical and material pollution can lead to toxic outcomes, that energy supply is less rigid than once seemed, and that population density and ecological stresses lead to catastrophic outcomes for the ecological network of which we are a part. These themes are inspiring researchers in all fields to think about how they can address sustainability and climate grand challenges using their disciplinary tools. The trouble is: the problems are not disciplinary *and* there are no physical rules to guide sustainable design. In the absence of such guidance, we risk repeating a pattern of innovation that so far has led to environmental damage. The best way forward is to leverage our collective knowledge of the environmental system and of functional material and process performance in a coordinated, data-driven effort at a scale. We must curate and cultivate *big data* to inform design in a way that bridges materials and environmental disciplines to accelerate the path to a sustainable future.

Status quo.

Today, for many fields of nanotechnology, there is no such thing as “big data.” Big data are defined as extremely large datasets (particularly with respect to volume and variability, but potentially lacking in veracity) that have some potential value but are beyond the reaches of human trend or pattern recognition. The reason for a lack of big data in nanotechnology is that experimental successes are hard fought, costly and time consuming. Further, due to strong interest in phenomenological demonstration, many findings are not replicated in the original study nor replicable in subsequent investigation. This paucity of systematic data collection in nanotechnology is particularly acute in environmental health and safety (EHS) research (Figure 1), as is often the case in material innovation.



Since the industrial revolution, the status quo in design optimization has been to focus squarely on performance and cost metrics to translate a product to market, at scale, as quickly as possible (Box 1). Nanomaterials were among the first where a concerted effort was made to consider environmental implications early in the design phase of the materials. The evidence of this exists as tangible research centers and initiatives established in the United States and Europe. Yet, in spite of this great advance, here we sit: approaching 20 years after those first calls to action were issued (circa 2003[1,2]) and with few emergent design rules to guide the sustainable development of nanomaterials. This is a consequence of a limited data collection guided by coordinated efforts between materials and environmental health and safety (EHS) researchers. In other words, the data are too few and too far apart.



This heterogeneous landscape of discovery is normal in any nascent field; there is a necessary induction period where research proceeds in multiple promising channels and best practices emerge. Moving beyond this phase, standardization of data reporting *could* create an opportunity to augment and accelerate subsequent breakthroughs. Several early nano EHS leaders saw such a need for in characterization practices[3]. Investigators quickly learned that nominal assignments (e.g., "nanotubes") referred to classes of materials instead of individual entities, in contrast to the high level of specification that is implied by the name of a chemical. As such, minimum characterization criteria (such as particle size, shape, composition, and purity) were defined to help answer the very basic question of *what, exactly, was the topic of study for a particular article*. These minimum criteria raised the bar for experimentalists, and the additional data helped researchers tie observations of EHS effects to some implicit characteristic of the materials. Nevertheless, a lack of standardization in reporting the data (e.g., systematic units, standard language, and the location of the data (i.e., main text versus supplemental documentation)) continues to limit the machine readability of contributions. Here, we emphasize that the qualitative and relative measures that are often part of nanomaterial characterization, such as electron micrographs or x-ray photoelectron spectra, were and are far from being standardized to a point of inter-article machine interpretability; this is just emerging for self-contained studies[4–6] and is not to a point of broad application that would aid interoperable study between many articles. This future is on our horizon[7,8].

Accelerating discovery.

Advances in computer science permeate every corner of society and will change all fields of science and engineering. The value of the experimentalist remains- and is perhaps accentuated- but these experimentalists are challenged to adapt to an environment where non-human colleagues will be reading their articles. Olivetti and colleagues[9] helped shepherd this idea in an effort to elucidate synthetic pathways by mining many thousands of contributions using natural language processing and striving to identify best chemistries via machine learning techniques (manual efforts in Green Chemistry pathways were conducted earlier[10]). While making great progress in the development of the approach and algorithm, limitations emerged that resulted directly from the *way* that we report data. Principally, (1) we lack routine in how we articulate

procedure, (2) we do not report synthetic failures, and (3) one of the only universally reported parameters is temperature, falling far short of the detail needed to reproduce an experiment (Box 1). The consequence of these choices is severe: new insights cannot be extracted from a collective body of literature when only a handful of studies have information presented in a findable, consistent manner.

The benefit of interoperable data is simply summarized: the whole should be greater than the sum of the parts. Recognizing this, a consortium of scientists proposed a set of standards that encourages researchers to report and archive data in a way that is findable, accessible, interoperable, and reusable (FAIR[11]), going so far as to encourage the "Group of 20" (G20) leaders to endorse it and make it part of a Nanotechnology Roadmap for 2030[12]. In the US, FAIR practices are now being required or strongly encouraged by several funding agencies, and that makes good sense: the public investment in the research enterprise should maximize its impact by extending beyond the benefits of the original investigation. The challenge is that individual principle investigators (PIs) working outside of fields that have standardized data repositories (i.e., most of nanotechnology) are often left to develop their own interpretation of FAIR practices. To combat this, some have skillfully defined best practices for collating metadata around procedure, minimum datasets, and even specifying units and tabular reporting conditions (see Chetwynd, Wheeler, and Lynch's 2019 study, *Best practice in reporting corona studies: Minimum information about Nanomaterial Biocorona Experiments (MINBE)* [13]; now supported on the Open Science Framework; osf.io). Importantly, the authors note that their proposed system is flexible; it is not meant to constrain creativity, discourage bold experimentation, or provide an exclusionary, proscribed, or preconceived definition of "good work." Instead, the point is to enable learning through intercomparable and reusable data. This is critically important in all fields, but especially in the very complex systems of nanoEHS, where both the material and receiving environments encompass a vast descriptor space. We are at a crossroads where we can no longer throw our hands up and say "it is too complex to understand," but instead must strive to generate the information needed to make sense of it all. Indeed, the type of challenge that is outside the reach of reductionist human comprehension is exactly what machine learning and big data are here to help us tackle.

Multi-objective optimization for sustainability.

The challenge of sustainable design is unique in this space because it requires systematic data collection across at least two disciplines, one that is focused on functional performance and one that is focused on the environmental performance (sometimes called "implications"). The reason for this seems obvious on its face: no amount of environmental optimization in material synthesis is going to be meaningful if the product is not competitive and profitable. Conversely, no material designer can hope to achieve sustainability goals if the guidelines for environmental design are not there. The same materials must be evaluated alongside one another for both metrics of success, and when they cannot be collected simultaneously, the data from independent studies must be interoperable so they can be analyzed as a collective later. Gilbertson et al. ([14]) described an elegant approach to this dual challenge, where the structure of a

material gives rise to a measurable property that can impart either linked or decoupled (i.e., independently tunable) functional and environmental performance metrics. Early work from our group and colleagues[15–17] illustrated a similar approach in synthetic chemical optimization, where the pairing of the two approaches yielded unique insights to the nanomaterial field. To date, these have been small data efforts, but the coordination between the materials and environmental teams established an important principle for effective discovery and multi-objective optimization.

Multi-objective optimization has a history in materials and mechanical engineering, and that history can be leveraged to guide the design of interoperable datasets. Specifically, Michael Ashby famously collated vast datasets of materials performance metrics (e.g., stiffness, strain, and density) and later started to include embodied energy (which can be related to greenhouse gas footprint if one knows the energy source) as an environmental design metric. This type of cross-disciplinary data collection is urgently needed in nanomaterial research and is only starting to emerge[18,9]. The urgency is underscored by the pace of materials innovation, where functional performance and cost metrics are always prioritized in order to maintain economic viability (a basic necessity in a young company's operation). If there is a hope for inclusion of environmental metrics in the design process, they have to be at-the-ready and poised to guide even untested chemistries. This can only happen with large volumes of data coming together to elucidate design *principles*. Thus, we encourage teams of researchers working on both sides of this problem to collect and systematically report parametric data surrounding environmental and functional performance objectives, such that fundamental mechanisms and drivers for each may be identified by human or machine.

Supporting a better future.

To facilitate this bold vision, several publishing groups are building critical supporting infrastructure, including encouraging statements on data availability, expanding their online data storage file formats and file size limitations, and even setting up consulting services to help authors identify the best options for publicly-available storage of a particular type of data (see recent commentary in *Nature Nanotechnology*[19] and other recent databases[7,20,21]). What remains is for subfields to define specific metrics, units, keywords (i.e., to aid machine readability), and, importantly, to report failures as much as successes to bound the design and synthesis space. Salient experimental details should always be included in the main text, but a standard set of experimental parameters and performance metrics that are systematically stored in the supplemental documentation (often available free of charge) would aid machine readability while preserving the flexibility required for creative communication of novel scientific insights. Example basic process metrics might include efficiencies of energy and atoms (such as atom economy, yield, and energy demand). Absent the ability to report these metrics, researchers can provide the raw information necessary for calculation, which would support retroactive evaluation of the process or material in a future where new or potentially unifying metrics are identified. Finally, EHS and materials researchers should coordinate the choice of study materials to ensure that paired environment and

functional performance datasets are available, whether collected in concert, series, or parallel, to support multi-objective optimization (Box 1).

If we truly want to realize the sustainable-by-design future we seek, both materials and environmental experimentalists must start making decisions to accelerate the pace of discovery. Experimentation and data preservation must be premeditated with a future of computational treatment in mind. Such an approach will not only serve to close the gaps between experimentation and computation, and between materials designers and environmental scientists, in order to meet the urgent need for a more resilient, sustainable society.

Competing Interests:
The authors declare no competing interests.

References

1. Commission, E. Towards a European strategy for nanotechnology. *Nanotechnol. Commun.* 1–28 (2004).
2. Dowling, a *et al.* Nanoscience and nanotechnologies : opportunities and uncertainties. *London R. Soc. R. Acad. Eng. Rep.* **46**, 618–618 (2004).
3. Grassian, V. H. Environmental Science: Nano-a journal is born: A new journal with a large scope that focuses on small materials. *Environ. Sci. Nano* **1**, 8–10 (2014).
4. Dee, N. T. In situ monitoring and control of carbon nanotube synthesis. (Massachusetts Institute of Technology, 2020).
5. ImageNet. (2020).
6. Kalinin, S. V. *et al.* Big, Deep, and Smart Data in Scanning Probe Microscopy. *ACS Nano* **10**, 9068–9086 (2016).
7. Yan, X., Sedykh, A., Wang, W., Yan, B. & Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* **11**, (2020).
8. Paunovska, K., Loughrey, D., Sago, C. D., Langer, R. & Dahlman, J. E. Using Large Datasets to Understand Nanotechnology. *Adv. Mater.* **31**, 1–16 (2019).
9. Kim, E. *et al.* Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **29**, 9436–9444 (2017).
10. Shi, W., Xue, K., Meshot, E. R. & Plata, D. L. The carbon nanotube formation parameter space: data mining and mechanistic understanding for efficient resource use. *Green Chem.* **19**, 3787–3800 (2017).
11. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
12. Haase, A. EU US Roadmap Nanoinformatics 2030. 0–2 (2017) doi:10.5281/zenodo.1486012.
13. Chetwynd, A. J., Wheeler, K. E. & Lynch, I. Best practice in reporting corona studies: Minimum information about Nanomaterial Biocorona Experiments (MINBE). *Nano Today* **28**, 100758 (2019).
14. Gilbertson, L. M., Zimmerman, J. B., Plata, D. L., Hutchison, J. E. & Anastas, P.

T. Designing nanomaterials to maximize performance and minimize undesirable implications guided by the Principles of Green Chemistry. *Chem. Soc. Rev.* **44**, 5758–5777 (2015).
15. Meshot, E. R. *et al.* Engineering Vertically Aligned Carbon Nanotube Growth by Decoupled Thermal Treatment of Precursor and Catalyst. *ACS Nano* **3**, 2477–2486 (2009).
16. Plata, D. L., Hart, A. J., Reddy, C. M. & Gschwend, P. M. Early Evaluation of Potential Environmental Impacts of Carbon Nanotube Synthesis by Chemical Vapor Deposition. *Environ. Sci. Technol.* **43**, 8367–8373 (2009).
17. Plata, D. L., Meshot, E. R., Reddy, C. M., Hart, A. J. & Gschwend, P. M. Multiple Alkynes React with Ethylene To Enhance Carbon Nanotube Synthesis, Suggesting a Polymerization-like Formation Mechanism. *ACS Nano* **4**, 7185–7192 (2010).
18. Falinski, M. M. *et al.* A framework for sustainable nanomaterial selection and design based on performance, hazard, and economic considerations. *Nat. Nanotechnol.* **13**, 708–714 (2018).
19. The importance and challenges of data sharing. *Nat. Nanotechnol.* **15**, 83 (2020).
20. Nano: A Nature Portfolio Solution. https://nano.nature.com.
21. Nanotechnology Products Database.

Figure 1. Disparities between material and chemical innovation and environmental health and safety (EHS) research reflected by the number of publication records. The number of contributions in EHS research is often orders of magnitude behind. (Data from Web of Science; accessed December 8, 2020; nanomaterial-related work only; carbon nanotube (CNT); nano-silver (nAg); nano-titania ($nTiO_2$)).